**The Journal of Animal & Plant Sciences, 26(3): 2016, Page: 594-598**
**ISSN: 1018-7081**

# EXAMINING OF MULTIPLE IMPUTATION METHOD IN TWO MISSING OBSERVATION MECHANISMS

Gazel Ser[1*] and Sıddık Keskin[2]

[1]Yüzüncü Yil University, Agriculture Faculty, Department of Animal Science, 65080 Van-Turkey
[2]Yüzüncü Yil University, Faculty of Medicine, Department of Biostatistics, 65080 Van-Turkey
[*]Corresponding author e-mail: gazelser@gmail.com

## ABSTRACT

This study contains an examination of the missing data structures, occurring in many fields, especially in livestock. It also examines the processes to obtain the solution for the missing data. For this purpose, linolenic acid measurements obtained from four different anatomic regions of two animal species were taken as dependent variables. For the dependent variable, the observations were deleted at the ratio of 10% and 20%, creating the missing structures of missing completely at random (MCAR) and missing at random (MAR). Subsequently, these data sets were completed using multiple imputation (MI) method. Generalized Estimating Equation (GEE) and mixed model methods were used in the missing data structures and for the purpose of evaluating the data completed with MI. The study were obtained almost same results obtained from GEE and mixed model in the missing data structures. At the same time, there was not found difference between the methods in completed data using MI method.

As a result, it is stated that valid results obtained in missing data structures by used GEE and mixed model analysis. When these results are also compared, it can be concluded that multiple imputation (with these ratio of missing) is not necessary before GEE and mixed model.

**Keywords:** Missing data analysis, repeated data, generalized estimating equation, mixed model

## INTRODUCTION

The concept of missing data means that observations of the variables examined in any study have not been obtained due to various reasons (Little and Rubin 1987). Throughout any research process, all measurements are sought to be as complete as possible. A study based on complete data is significant and allows one to draw more reliable conclusions. It is also important to determine the number of missing observations and the reasons for their absence (Tabachnick and Fidell 2007). For example, in research conducted on livestock, some observation values may not be obtained for the variables interested due to the mortality or morbidity of animals, or problems related to the researcher.

Standard statistical methods such as ANOVA and MANOVA assume that a data set is in rectangular form. The analysis are performed by assuming that variables are presented in columns and the measurements of these variables are presented in rows in this rectangle. Thus, in the event that there are missing observations in a study, this rectangle’s shape is deformed and standard statistical methods cannot be applied (Little and Rubin 1987). According to Schafer and Graham (2002), rather than completely removing individuals or units from the analysis, it would be more proper to perform analysis over completed data sets by imputing values to missing observations using information obtained from the units with missing data. At this point, researchers prefer two ways. The first approach comprises deleting methods such as listwise and pairwise deletion. However, as processes are executed over observations with complete data in these methods, missing data are neglected. The second approach comprises value imputation methods, wherein missing values are estimated and these methods are divided into two: single and multiple imputation (MI). As only one imputed value is calculated for each missing value in the single imputation method, variability originating from missing data is ignored. The MI method imputes different values for missing observations within the imputation process. As the method considers the uncertainty from the imputation of missing values, it provides a significant advantage for researchers by comparison with other imputation methods. Moreover, MI provides more valid estimations, such as standard error, regression coefficient, and average, compared to other value imputation methods (Twisk and de Vente 2002; Fichman and Cummings 2003).

Although the studies on methods used in missing observation analysis are numerous in fields such as the social, behavioral, medical, and educational sciences, they are not common in the field of agriculture. For this purpose, by creating two missing observation mechanisms (MAR and MCAR), it was aimed to evaluate performance of the multiple imputation method, which has been commonly used by the researchers in the recent years, in the Generalized Estimating Equations (GEE) and mixed model.

## MATERIALS and METHODS

**Animal Sampling:** Herein, the data set used as the application material was obtained from 47 heads of male lambs and kids. The data comprised the measurements of linolenic acid (C18:3) content from four different anatomic regions (*subcutaneous fat* (SF), *Longissimus dorsi* (LD), *Semimembranosus* (SM), and *Triceps brachii* (TB)) of slaughtered animals that fattened under two different feeding system: concentrate and pasture.

**Creation of the Missing Data Mechanism:** Recently, the missing data mechanisms appearing in scientific studies come in the forms of missing completely at random (MCAR), which is a structure independent of both observed and non-observed data; missing at random (MAR), which is a structure dependent only on observed data; and missing not at random (MNAR), which is a structure dependent on both observed and non-observed data. These structures were first classified by Rubin (1976). In the analysis of missing data structures, both quantitative and qualitative information are obtained on missing data (Maroto-Molina *et al*. 2013).

This study begins with a complete data set containing no missing observations. Subsequently, some observations over this data set are deleted at certain ratios (10% and 20%) and two structures (MCAR and MAR) of missing observations are created. Thus, five data sets were used in the study. Because the MCAR structure is independent of observed and non-observed values, this structure was created by randomly deleting observations of the dependent variable (linolenic acid) at 10% and 20% rates. The suitability of the data sets with the MCAR structure was tested by applying Little's MCAR test. As the MAR structure is dependent on observed data, the animals having the highest values in the dependent variable, in the t = t anatomic region were assumed as missing, at the same deletion rates, from the latter, t = t + 1 anatomic region (Twisk and de Vente 2002). Therefore, the measurements in the t = 1 region in the data set with the MAR structure were obtained as the full data set. Deleting was not performed on independent variables included in model.

**MI Method:** The MI method has been the primary value imputation method most frequently used in recent studies. This method comprises three phases: imputation, analysis, and pooling. In the first phase, missing observations in the data set are imputed to the appropriate values *m* times (the number of imputation in this study is determined as *m*=5). In the second phase, parameter estimations and standard errors are obtained from *m* completed data sets. Statistical results are assembled from the *m* completed data sets combined in the third phase. The combination of the uncertainty in the parameters of these *m* complete data sets has been expressed by two variance components suggested by Rubin (1987). The first component is within-imputation variance and the second is between-imputation variance. Between-imputation variance is accepted as a criterion of the uncertainty arising from missing data (Twisk and de Vente 2002; Baraldi and Enders 2010; Twisk *et al*. 2013).

**Statistical Analysis:** The analysis methods of GEE and mixed model used in repeated measurement structures were preferred in the study for examination of the missing observation structures and evaluation of multiple imputation. The majority of standard statistical methods (e.g., chi-square and linear regression) used in analysis of data are based on the assumption that variables are independent of one another. This assumption will not be valid for observations having a repeated structure. The reason is the measurements of the same unit will tend to be correlated with those taken successively. Thus, methods that consider this correlation structure are preferred. GEE and mixed model are the primary ones among these methods. In GEE method, the structure of the relationship between repeated measurements is defined using working correlation matrix, whereas in mixed models, it is considered using variance-covariance structures (Yazıcı 2001; Fitzmaurice *et al* 2004; Aktaş 2005; Ser 2011). In GEE analysis, the exchangeable working correlation structure assumed that the correlation between any pair of measurements on the same individual is the same. In addition, compound symmetry structure based on homogenous variance and covariance between measurements was used in mixed model application. The study was conducted in two phases. In the first phase, observations were deleted at the ratio of 10% and 20% on the dependent continuous variable to create MAR and MCAR data sets. Parameter estimations were obtained using GEE and mixed model methods. In the second phase, missing observations were estimated using MI method in the data sets with MAR and MCAR structures. Parameter estimations were obtained by applying GEE and mixed analysis again to the data sets completed. In the study, GEE, mixed model and MI analysis were performed using SAS procedures, PROC GENMOD, PROC MIXED and PROC MI, respectively (SAS 2014).

## RESULTS and DISCUSSION

In Table 1, descriptive statistics of the complete data set and the MCAR and MAR structures are provided.

**Table 1. Descriptive statistics for the dependent variable**

| | N | Mean | Std. Error |
|---|---|---|---|
| Complete Data | 188 | 1.079 | 0.056 |
| | **Missing at Random (MAR) Mechanism** | | |
| 10% Missing Data | 169 | 1.122 | 0.060 |
| Multiple Imputation | 188 | 1.095 | 0.055 |
| 20% Missing Data | 150 | 1.183 | 0.066 |
| Multiple Imputation | 188 | 1.105 | 0.055 |
| | **Missing Comletely at Random (MCAR) Mechanism** | | |
| 10% Missing Data | 169 | 1.071 | 0.058 |
| Multiple Imputation | 188 | 1.080 | 0.054 |
| 20% Missing Data | 150 | 1.027 | 0.061 |
| Multiple Imputation | 188 | 1.073 | 0.053 |

In Table 1, the number of observations (N), mean and standard error, are provided. The standard error estimations of MI, for both deletion rates on the MCAR structured data sets are lower than those of the complete data set. However, the estimations for both rates on the MAR structures are close to those on the complete data set. In Table 2 gives the results of the GEE and mixed model analysis applied to the data sets with MAR and MCAR structures, which were created by deletion of the observations at the ratios of 10% and 20% from complete data.

**Table 2. Results for GEE and mixed model analysis**

| | | | **GEE Results** | | | |
|---|---|---|---|---|---|---|
| | | **Species** | **Feeding System** | | **Anatomic[1] Regions** | |
| | **Intercept** | **Male lambs** | **Concentrate** | **SF** | **LD** | **SM** |
| Complete Data | 1.883 | 0.486 | -1.387 | -0.610 | -0.203 | -0.084 |
| | (0.054)$^{***}$ | (0.055)$^{***}$ | (0.055)$^{***}$ | (0.051)$^{***}$ | (0.051)$^{***}$ | (0.051)$^{ns}$ |
| MAR- 10% | 1.881 | 0.495 | -1.381 | -0.621 | -0.187 | -0.090 |
| | (0.055)$^{***}$ | (0.056)$^{***}$ | (0.054)$^{***}$ | (0.056)$^{***}$ | (0.056)$^{**}$ | (0.056)$^{ns}$ |
| MAR- 20% | 1.885 | 0.551 | -1.345 | -0.663 | -0.202 | -0.117 |
| | (0.056)$^{***}$ | (0.058)$^{***}$ | (0.055)$^{***}$ | (0.061)$^{***}$ | (0.061)$^{**}$ | (0.061)$^{ns}$ |
| MCAR- 10% | 1.899 | 0.492 | -1.356 | -0.647 | -0.232 | -0.125 |
| | (0.053)$^{***}$ | (0.053)$^{***}$ | (0.053)$^{***}$ | (0.054)$^{***}$ | (0.058)$^{**}$ | (0.056)$^{*}$ |
| MCAR- 20% | 1.879 | 0.456 | -1.338 | -0.621 | -0.207 | -0.142 |
| | (0.056)$^{***}$ | (0.053)$^{***}$ | (0.053)$^{***}$ | (0.059)$^{***}$ | (0.066)$^{**}$ | (0.064)$^{*}$ |
| | | | **Mixed Model Results** | | | |
| Complete Data | 1.883 | 0.486 | -1.387 | -0.610 | -0.203 | -0.084 |
| | (0.054)$^{***}$ | (0.056)$^{***}$ | (0.056)$^{***}$ | (0.051)$^{***}$ | (0.051)$^{***}$ | (0.051)$^{ns}$ |
| MAR- 10% | 1.881 | 0.495 | -1.382 | -0.622 | -0.187 | -0.090 |
| | (0.055)$^{***}$ | (0.057)$^{***}$ | (0.056)$^{***}$ | (0.055)$^{***}$ | (0.055)$^{**}$ | (0.055)$^{ns}$ |
| MAR- 20% | 1.885 | 0.551 | -1.345 | -0.663 | -0.202 | -0.117 |
| | (0.056)$^{***}$ | (0.060)$^{***}$ | (0.056)$^{***}$ | (0.061)$^{***}$ | (0.061)$^{**}$ | (0.061)$^{ns}$ |
| MCAR-10% | 1.899 | 0.491 | -1.357 | -0.647 | -0.232 | -0.125 |
| | (0.054)$^{***}$ | (0.054)$^{***}$ | (0.054)$^{***}$ | (0.054)$^{***}$ | (0.055)$^{***}$ | (0.055)$^{*}$ |
| MCAR- 20% | 1.879 | 0.456 | -1.339 | -0.621 | -0.207 | -0.142 |
| | (0.056)$^{***}$ | (0.053)$^{***}$ | (0.053)$^{***}$ | (0.058)$^{***}$ | (0.065)$^{**}$ | (0.064)$^{*}$ |

Reference parameter: Kids, Pasture and *Triceps brachii* (TB); [1]*:Subcutaneous fat* (SF), *Longissimus dorsi* (LD), *Semimembranosus*(SM); $^{*}$: p<0.05; $^{**}$: p<0.01; $^{***}$: p<0.001; $^{ns}$: non-significant

In Table 2, the regression coefficients and standard errors obtained from GEE and mixed model analysis applied to MAR, MCAR and complete data sets for the continuous result variable (linoleic acid) are almost same. In addition, it was determined both missing observation structures had no effect on results of the model applied. Because both regression coefficients and standard errors came up with similar results in GEE and

Mixed model for MAR and MCAR structures. These results show similarity to those in the study conducted by Twisk (2004). In the study, the author has stated that GEE and random coefficient models applied for the continuous result variable gave valid results in incompleted data sets and that missing observation structures had no effect on results of the model. Nevertheless, the studies conducted often show that missing observation structures are determinant in model selections. For example, it is stated that, if the missing observation structure in the data is MAR, Weighted-GEE (WGEE) method provides valid and consistent parameter estimations, yet, in the event missing observation structure is MCAR, GEE method provides valid parameter estimations (Molenbergs and Kenward 2007; Sun 2013).

Missing observations were completed through MI method in MAR and MCAR data sets and GEE and mixed model was applied again, and the results obtained are given in Table 3.

**Table 3. Results for GEE and mixed model analysis related to MI estimations**

| | | | **GEE Results** | | | |
|---|---|---|---|---|---|---|
| | | **Species** | **Feeding System** | | **Anatomic[1] Regions** | |
| | **Intercept** | **Male lambs** | **Concentrate** | **SF** | **LD** | **SM** |
| Complete Data | 1.883 | 0.486 | -1.387 | -0.610 | -0.203 | -0.084 |
| | (0.054)$^{***}$ | (0.055)$^{***}$ | (0.055)$^{***}$ | (0.051)$^{***}$ | (0.051)$^{***}$ | (0.051)$^{ns}$ |
| MAR-MI (10%) | 1.866 | 0.516 | -1.368 | -0.612 | -0.178 | -0.074 |
| | (0.052)$^{***}$ | (0.051)$^{***}$ | (0.051)$^{***}$ | (0.054)$^{***}$ | (0.054)$^{*}$ | (0.054)$^{ns}$ |
| MAR-MI (20%) | 1.889 | 0.536 | -1.355 | -0.646 | -0.209 | -0.128 |
| | (0.051)$^{***}$ | (0.049)$^{***}$ | (0.049)$^{***}$ | (0.053)$^{***}$ | (0.053)$^{***}$ | (0.053)$^{*}$ |
| MCAR-MI (10%) | 1.903 | 0.472 | -1.348 | -0.647 | -0.262 | -0.130 |
| | (0.052)$^{***}$ | (0.051)$^{***}$ | (0.051)$^{***}$ | (0.052)$^{***}$ | (0.052)$^{***}$ | (0.052)$^{*}$ |
| MCAR-MI (20%) | 1.876 | 0.472 | -1.329 | -0.631 | -0.215 | -0.159 |
| | (0.048)$^{***}$ | (0.044)$^{***}$ | (0.045)$^{***}$ | (0.054)$^{***}$ | (0.054)$^{***}$ | (0.054)$^{*}$ |
| | | | **Mixed Model Results** | | | |
| Complete Data | 1.883 | 0.486 | -1.387 | -0.610 | -0.203 | -0.084 |
| | (0.054)$^{***}$ | (0.056)$^{***}$ | (0.056)$^{***}$ | (0.051)$^{***}$ | (0.051)$^{***}$ | (0.051)$^{ns}$ |
| MAR-MI (10%) | 1.866 | 0.516 | -1.368 | -0.612 | -0.178 | -0.074 |
| | (0.053)$^{***}$ | (0.052)$^{***}$ | (0.052)$^{***}$ | (0.054)$^{***}$ | (0.054)$^{**}$ | (0.054)$^{ns}$ |
| MAR-MI (20%) | 1.889 | 0.536 | -1.355 | -0.646 | -0.209 | -0.128 |
| | (0.051)$^{***}$ | (0.050)$^{***}$ | (0.051)$^{***}$ | (0.052)$^{***}$ | (0.052)$^{***}$ | (0.052)$^{*}$ |
| MCAR-MI (10%) | 1.903 | 0.472 | -1.348 | -0.647 | -0.262 | -0.129 |
| | (0.052)$^{***}$ | (0.052)$^{***}$ | (0.052)$^{***}$ | (0.052)$^{***}$ | (0.052)$^{***}$ | (0.052)$^{*}$ |
| MCAR-MI (20%) | 1.876 | 0.473 | -1.329 | -0.631 | -0.215 | -0.159 |
| | (0.049)$^{***}$ | (0.045)$^{***}$ | (0.045)$^{***}$ | (0.054)$^{***}$ | (0.054)$^{***}$ | (0.054)$^{*}$ |

Reference parameter: Kids, Pasture and *Triceps brachii* (TB); [1]:*Subcutaneous fat* (SF), *Longissimus dorsi* (LD), *Semimembranosus*(SM); $^{*}$: p<0.05; $^{**}$: p<0.01; $^{***}$: p<0.001; $^{ns}$: non-significant

In Table 3, regression coefficients obtained from the GEE and mixed model analysis applied to MI estimations, yet, with slight differences in the standard error. When the complete data sets analyzed through both models were compared with MI data sets, quite varied regression coefficients were obtained. In addition, standard errors were obtained partially bigger compared to complete data in the anatomic regions included in the model as variable, but smaller for other categorical independent variables. Hence, MAR and MCAR structures were determined to demonstrate similar behaviors in the models used in evaluation of MI estimations. There are different approaches in the literature with regard to these results. For example, the study conducted by Twisk and de Vente (2002) has found that, in MCAR and MAR structures, regression coefficients and standard errors were bigger in MANOVA and GEE models compared to complete data. In addition, Baraldi and Enders (2010) stated that MI method was applicable in both structures (MAR or MCAR), and Allison (2002) and McKnight *et al* (2007) emphasize that consistent estimations with small deviations could be obtained when MI method was applied under MAR assumption.

**Conclusion:** The parameter estimations obtained from GEE and mixed model analysis applied to the data sets with MAR and MCAR structures in the first phase of this study, which was conducted in two phases, were too similar. Results of GEE and mixed model obtained in the data sets completed through MI method in the second phase of the study were too similar to each other. Hence, no difference was found between two models in missing data and completed data sets. Therefore, it is possible to say that missing observation estimation is unnecessary at these missing observation ratios.

**Acknowledgement:** The authors would like to thank Professor Aşkın KOR and Assist. Prof. Serhat KARACA for providing data set.

## REFERENCES

Aktaş, A. (2005). Generalized estimating equations (“GEE”). M.Sc. thesis, University of Hacettepe, Ankara, Turkey.

Allison, P. D. (2002). Missing data. Sage Publications, Inc., California.

Baraldi, A. N. and C. K. Enders (2010). An introduction to modern missing data analysis. J. School Psychol. 48: 5–37.

Fichman, M. and J. M. Cummings (2003). Multiple imputation for missing data: Making the most the most of what you know. Organ. Res. Methods 6: 282- 308.

Fitzmaurice, G.M., N.M. Laird and J. H. Ware (2004). Applied longitudinal analysis. John Wiley & Sons, Inc., New Jersey.

Little, J. R. and D. Rubin (1987). Statistical analysis with missing data. John Wiley & Sons, Inc., New York.

Maroto-Molina, F., A. Gómez-Cabrera, J. E. Guerrero-Ginel, A. Garrido-Varo, D. Sauvant, G. Tran,V. Heuzé and D.C. Pérez-Marín (2013). Handling of missing data to improve the mining of large feed databases. J.Anim. Sci. 91: 491–500.

McKnight, P. E., K. M. McKnight, S. Sidani and A. J. Figueredo (2007). Missing data: A gentle introduction. The Guilford Press, New York.

Molenberghs, G. and M. G. Kenward (2007). Missing data in clinical studies**.** John Wiley&Sons, Ltd., England.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. John Wiley & Sons, New York.

Rubin, D. B. (1976). Inference and missing data. Biometrika 63: 581-592.

SAS (2014). SAS/STAT. Statistical analysis system for Windows. Relase 9.4. SAS Institute Inc.

Schafer, J.L. and J. W. Graham (2002). Missing data: Our view of the state of the art. Psychol. Methods 7: 147–177.

Ser, G. (2011). Model selection and comparing optimization techniques in marginal and non-marginal multilevel generalized linear mixed model using missing observed longitudinal data. Ph.D thesis, University of Yuzuncu Yil, Van, Turkey.

Sun, A. (2013). Applying eeighted GEE for missing data analysis and sample size estimation in repeated measurement studies with dropout. PhD. thesis, University of Maryland,USA.

Tabachnik, B.G. and L. S. Fidell (2007). Using multivariate statistics. Pearson Education Inc., Boston.

Twisk, J. and W. de Vente (2002). Attrition in longitudinal studies: How to deal with missing data. J. Clin. Epidemiol. 55: 329–337.

Twisk, J., M. de Boer, W. de Vente and M. Heymans (2013). Multiple imputation of missing values was not necessary before performing a longitudinal mixed-model analysis. J. Clin. Epidemiol. 66: 1022-1028.

Twisk, J.W.R. (2004). Longitudinal data analysis. A comparison between generalized estimating equations and random coefficient analysis. European J. Epidemiology 19: 769-776.

Yazıcı, B. (2001). Generalized estimating equation approach in the case of existence a covariate in categorical data analysis and an application. Ph.D thesis, University of Anadolu, Eskişehir, Turkey.